HANDHELD DEVICE FOR SUBSTITUTION FROM VISION TO AUDITION

Jean Rouat, Damien Lescal and Sean Wood

Neuroscience and Intelligent Signal Processing Research group
Département de génie électrique et génie informatique, université de Sherbrooke
2500 boul. de l'université, Sherbrooke, Québec, Canada, J1K 2R1
[jean.rouat,damien.lescal,sean.wood]@usherbrooke.ca

ABSTRACT

Sensorial substitution has great potential in rehabilitation, education, games, and in the creation of music and art. Current technologies allow us to develop sensorial substitution and sonification systems that would not have been imaginable two decades ago. It is desirable to let a large audience use and test sonification systems to provide feedback and improve their design. Handheld devices like smartphones or tablets include network connectivity (WIFI and/or Cellular radio) that can be used to transmit anonymous information about the configuration and strategies adopted by users. It is now feasible to obtain feedback from any user of substitution and sonification technology and not only from a limited number of subjects in the laboratory. Testing in the field with a large number of users is now possible thanks to telecommunication networks and machine learning tools to analyze big data. This work presents a handheld implementation of a simple video sonification system designed to test the acceptability of vision to audition substitution systems and in the near future to provide feedback from users. A first beta version was publicly released in November 2013 as an iOS application for large scale testing. The extended abstract introduces the interface and the underlying technology.

1. INTRODUCTION

Most sonification systems encode raw low resolution images into series of sounds that characterize the full scene and let the user interpret the visual scene based on acoustical cues and sounds. For example, the vOICe [2] system scans the image from left to right and encodes the mean pixel value as the volume, and the vertical position as the frequency component of the sound. The complete image is encoded regardless of the image content. A similar approach is often used in music composition and creation systems.

The proposed portable implementation has a different approach. In an effort to avoid overloading the user's audition, the system encodes salient features of the video stream based on contours, contrasts, and textures. Different musical notes - depending on the position of the most salient area - are played and positioned in a spatial auditory scene. It can therefore be used to locate objects or as a toy to compose simple musical melodies by moving the device around a simple object.

2. 3D VIRTUAL SOUND SPATIALIZATION

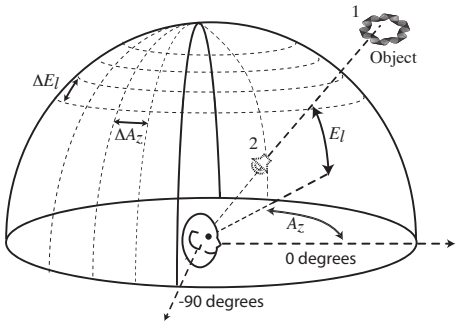


Figure 1: Synthesizing a virtual sound grid for auditory scene synthesis: number 1 represents a salient area of the image to be sonified, number 2 is the equivalent virtual sound source. The illusion of a 3D sound source at azimuth A_z and elevation E_l is created by convolving a monophonic sound with a pair of binaural filters stored in a matrix of Head Related Transfer Functions (HRTF) at the index corresponding to A_z and E_l. The depth (distance between the virtual source and the listener) is not yet encoded in this version of the system. The sound is a musical note.

An image is interpreted as a scene which is projected onto a sphere and characterized with spherical coordinates. The origin of the coordinate system is the center of the image. The horizontal axis of the image is mapped to azimuth, and the vertical axis to elevation (Figure 1). Areas of interest in the images are first found by a neural network [4]. The coordinates of these areas are then converted into their equivalent pairs (A_z, E_l).

Monophonic sounds are then spatialized by convolving with the Head Related Transfer Function (HRTF) filters corresponding to the (A_z, E_l) pairs. Theses measures are represented by a set of filters (Head Related Transfer Functions - HRTF) whose indexes are associated with A_z and E_l, [3]. For each (A_z, E_l) pair, a binaural filter can be retrieved from that HRTF and used to spatialize a monophonic sound. As pair of headphones are used, binaural (left and right) filtering is sufficient.

3. THE MAPPING PLATFORM

Conventional image processing requires the use of different filters to extract simultaneous contours and textures. This is not the case in this implementation. Salient parts of the images are found by a recurrent neural network that was designed for fast

computation and to simultaneously enhance gradients, contrasts, and textures in images [4]. Each pixel from the camera images are quantized into 256 grayscale levels and then associated with one neuron. Synapses connect each neuron to its 8 neighbors. Connecting weights w_{ij} between neurons n_i and n_j are bidirectional and given by: $w_{ij} = w_{ji} = f(|p_i - p_j|)$ where p_i is the pixel grayscaled value of neuron n_i. $|\ |$ is the absolute value and $f()$ is a user defined function that elicits strong connections between neurons having comparable pixel values. The update of each neuron is done by cumulating the state of the neighboring neurons while iterating. At iteration k: $s_i[k] = \frac{s_i[k-1]+\sum_j w_{ij}.s_j[k-1]}{9}$, where $s_i[k]$ is neuron n_i's state at iteration k, with $k < 5$. A salient point in the image corresponds to a neuron whose $s_i[k_{final}]$ is small. Neurons in highly textured areas and on contours are in fact weakly connected. On an iPhone or iPad, the averaged processing time, is on the order of 10 ms – 20 ms for one image.

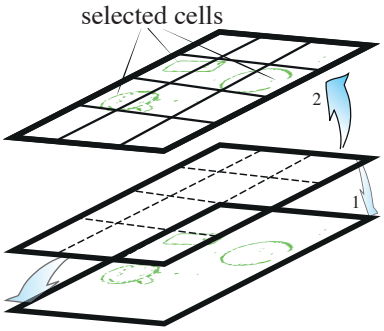
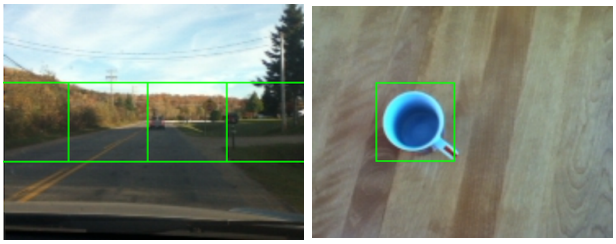


Figure 2: A grid comprising 3 x 4 cells is superimposed on the neural network outputs. The cells with the greatest number of salient points are selected and their attributed sounds are spatialized.

The number of salient points is computed for each cell in a 3x4 grid superimposed on the neural network (Figure 2). If the number of salient points for a given cell exceeds a predetermined threshold, then the cell from the grid is declared to be associated with an area of interest in the original image. The threshold is determined by the size of the handheld screen.



(a) On the road (b) Cup

Figure 3: Examples of area of interests found by the handheld device. (a): Four different notes are simultaneously played, with an elevation of 0^o and 4 azimuths of respectively 90, 30, -30 and -90 degrees. (b): One note is played with an azimuth of 30^o and an elevation of 0^o.

The positions of the 12 cells - from the 3D sound grid - were associated with 12 virtual sound sources located respectively at different azimuths A_z and elevations E_l (Figure 1). 4 azimuth angles (+90, 30, -30 and -90 degrees) and 3 elevation angles (-40, 0 and 45 degrees) were mapped respectively to each virtual source.

A *Virtual Acoustic Space* [5] is created through the headphones by convolving the user-created monophonic sounds with an Head Related Transfert Function (HRTF) filter before playing through the headsets. The 3D illusion is obtained by convolving the monophonic sound with the HRTF filters corresponding to the azimuth and elevation of the virtual sound source. The 12 cells of the grid were also mapped to notes from the C major musical scale so that small melody can be played by moving the device around an object (for example around a coffee mug).

4. CONCLUSION

A beta version [1] was publicly released in November 2013 as an iOS application and is available to the public for large scale testing.

As opposed to most existing sonification approaches, only salient parts of images are mapped to sounds to alleviate the overload of the hearing system and make the auditory scene easier to interpret. Spatial localisation and musical notes were combined for sound generation so that the same system can be used for navigation and simple melody generation.

Preliminary tests done with 10 blindfolded subjects showed that one can find objects and navigate in simple environments. Subjects compensated for the lack of elevation accuracy (as – in this version – the HRTF is not specific to the user) by moving the device up and down. Preliminary feedbacks showed that the simultaneous presentation of musical notes and spatialization was troublesome to some users. Therefore, two other versions will be developed, one specific to navigation (for blind people and serious gaming) and another one for melodic training and potential rehabilitation applications. In fact, to play a specific melody with the device, precise movements and displacements around an object are required. It implies good coordination skills of vision, audition and hand movements.

5. ACKNOWLEDGEMENT

Financial support from Québec FRNTQ and Univ. de Sherbrooke; the 3 anonymous reviewers for their constructive comments.

6. REFERENCES

[1] "See differently from vision to audition," http://www.gel.usherbrooke.ca/necotis/en/APP.html [Online].

[2] P. Meijer, "An experimental system for auditory image representations," *Biomedical Engineering, IEEE Transactions on*, vol. 39, no. 2, pp. 112 –121, 1992.

[3] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, 2001.

[4] D. Lescal, L.-C. Caron, and J. Rouat, "Neural visual objects enhancement for sensorial substitution from vision to audition," in *IEEE int. Conf. on Info. Sc., Signal Processing and their Applications*, 2012.

[5] J. Schnupp, I. Nelken, and A. King, *Auditory Neuroscience: Making Sense of Sound*. The MIT Press, 2011.